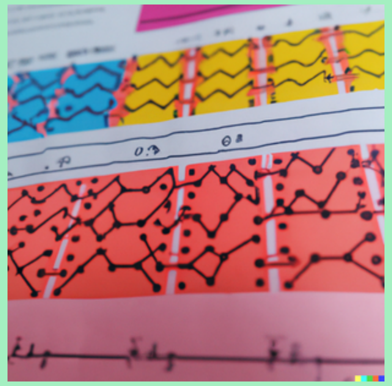
CASIMIR : un Corpus d’Articles Scientifiques Intégrant les ModIfications et Révisions des auteurs

Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez
{prénom.nom}@univ-nantes.fr



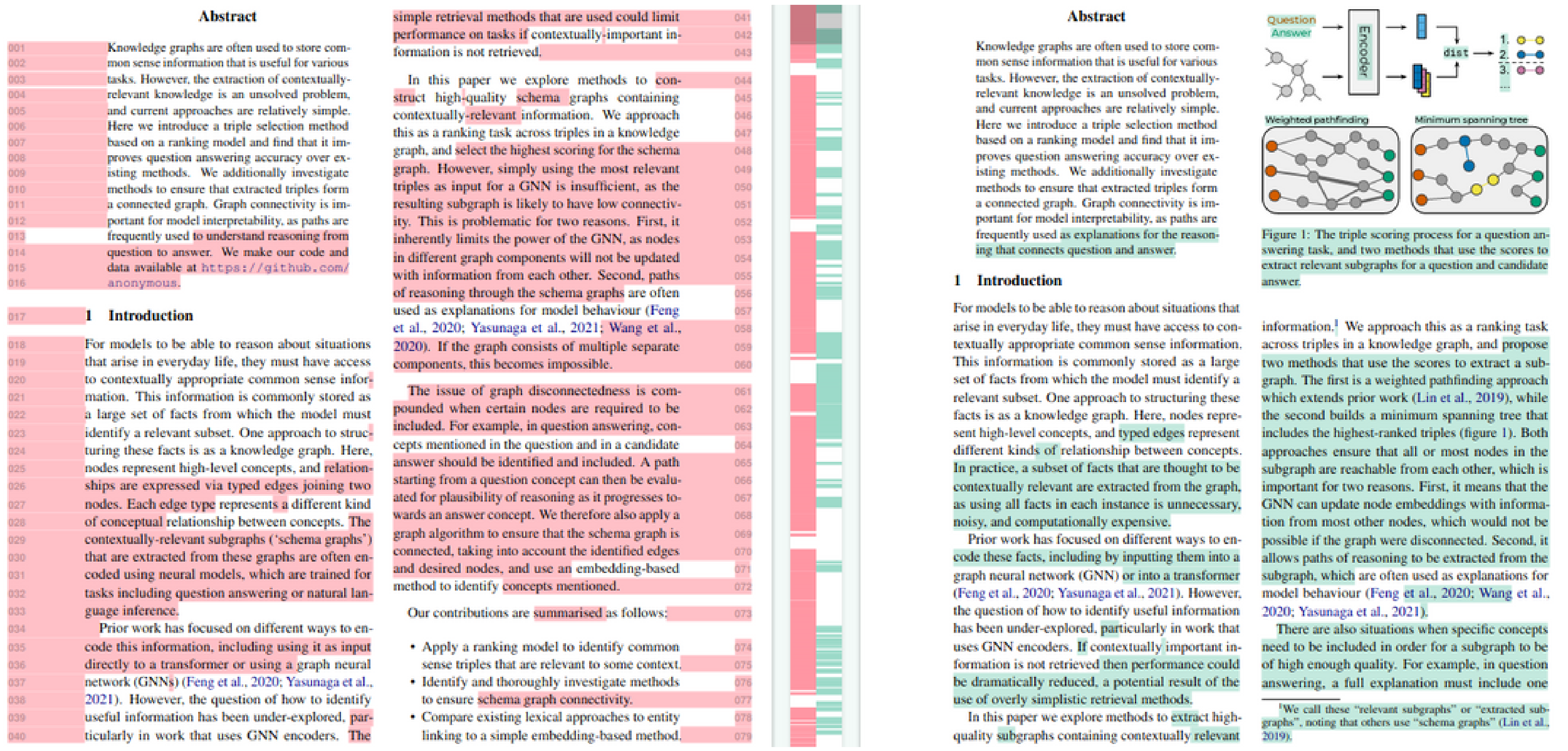
CONTEXTE

Motivations

- Écrire un article scientifique est une tâche difficile
- De bonnes compétences rédactionnelles sont indispensables
- Particulièrement pour les jeunes chercheurs et les non natifs anglophones

Objectif

Offrir une ressource sur la révision d'articles scientifiques



Abstract

Knowledge graphs are often used to store common sense information that is useful for various tasks. However, the extraction of contextually-relevant knowledge is an unsolved problem, and current approaches are relatively simple. Here we introduce a triple selection method based on a ranking model and find that it improves question answering accuracy over existing methods. We additionally investigate methods to ensure that extracted triples form a connected graph. Graph connectivity is important for model interpretability, as paths are frequently used to understand reasoning from question to answer. We make our code and data available at https://github.com/anonymous.

1 Introduction

For models to be able to reason about situations that arise in everyday life, they must have access to contextually appropriate common sense information. This information is commonly stored as a large set of facts from which the model must identify a relevant subset. One approach to structuring these facts is as a knowledge graph. Here, nodes represent high-level concepts, and relationships are expressed via typed edges joining two nodes. Each edge type represents a different kind of conceptual relationship between concepts. The contextually-relevant subgraphs ('schema graphs') that are extracted from these graphs are often encoded using neural models, which are trained for tasks including question answering or natural language inference.

Prior work has focused on different ways to encode this information, including using it as input directly to a transformer or using a graph neural network (GNNs) (Feng et al., 2020; Yasunaga et al., 2021). However, the question of how to identify useful information has been under-explored, particularly in work that uses GNN encoders. The simple retrieval methods that are used could limit performance on tasks if contextually-important information is not retrieved.

In this paper we explore methods to construct high-quality schema graphs containing contextually-relevant information. We approach this as a ranking task across triples in a knowledge graph, and select the highest scoring for the schema graph. However, simply using the most relevant triples as input for a GNN is insufficient, as the resulting subgraph is likely to have low connectivity. This is problematic for two reasons. First, it inherently limits the power of the GNN, as nodes in different graph components will not be updated with information from each other. Second, paths of reasoning through the schema graphs are often used as explanations for model behaviour (Feng et al., 2020; Yasunaga et al., 2021; Wang et al., 2020). If the graph consists of multiple separate components, this becomes impossible.

The issue of graph disconnectedness is compounded when certain nodes are required to be included. For example, in question answering, concepts mentioned in the question and in a candidate answer should be identified and included. A path starting from a question concept can then be evaluated for plausibility of reasoning as it progresses towards an answer concept. We therefore also apply a graph algorithm to ensure that the schema graph is connected, taking into account the identified edges and desired nodes, and use an embedding-based method to identify concepts mentioned.

Our contributions are summarised as follows:

- Apply a ranking model to identify common sense triples that are relevant to some context.
- Identify and thoroughly investigate methods to ensure schema graph connectivity.
- Compare existing lexical approaches to entity linking to a simple embedding-based method.

Abstract

Knowledge graphs are often used to store common sense information that is useful for various tasks. However, the extraction of contextually-relevant knowledge is an unsolved problem, and current approaches are relatively simple. Here we introduce a triple selection method based on a ranking model and find that it improves question answering accuracy over existing methods. We additionally investigate methods to ensure that extracted triples form a connected graph. Graph connectivity is important for model interpretability, as paths are frequently used as explanations for the reasoning that connects question and answer.

1 Introduction

For models to be able to reason about situations that arise in everyday life, they must have access to contextually appropriate common sense information. This information is commonly stored as a large set of facts from which the model must identify a relevant subset. One approach to structuring these facts is as a knowledge graph. Here, nodes represent high-level concepts, and typed edges represent different kinds of relationship between concepts. In practice, a subset of facts that are thought to be contextually relevant are extracted from the graph, as using all facts in each instance is unnecessary, noisy, and computationally expensive.

Prior work has focused on different ways to encode these facts, including by inputting them into a graph neural network (GNN) or into a transformer (Feng et al., 2020; Yasunaga et al., 2021). However, the question of how to identify useful information has been under-explored, particularly in work that uses GNN encoders. If contextually important information is not retrieved then performance could be dramatically reduced, a potential result of the use of overly simplistic retrieval methods.

In this paper we explore methods to extract high-quality subgraphs containing contextually relevant information.[1] We approach this as a ranking task across triples in a knowledge graph, and propose two methods that use the scores to extract a subgraph. The first is a weighted pathfinding approach which extends prior work (Lin et al., 2019), while the second builds a minimum spanning tree that includes the highest-ranked triples (figure 1). Both approaches ensure that all or most nodes in the subgraph are reachable from each other, which is important for two reasons. First, it means that the GNN can update node embeddings with information from most other nodes, which would not be possible if the graph were disconnected. Second, it allows paths of reasoning to be extracted from the subgraph, which are often used as explanations for model behaviour (Feng et al., 2020; Wang et al., 2020; Yasunaga et al., 2021).

There are also situations when specific concepts need to be included in order for a subgraph to be of high enough quality. For example, in question answering, a full explanation must include one

Figure 1: The triple scoring process for a question answering task, and two methods that use the scores to extract relevant subgraphs for a question and candidate answer.

[1] We call these "relevant subgraphs" or "extracted subgraphs", noting that others use "schema graphs" (Lin et al., 2019).

Exemple de révisions

CASIMIR : un Corpus d'Articles Scientifiques Intégrant les ModIfications et Révisions des auteurs

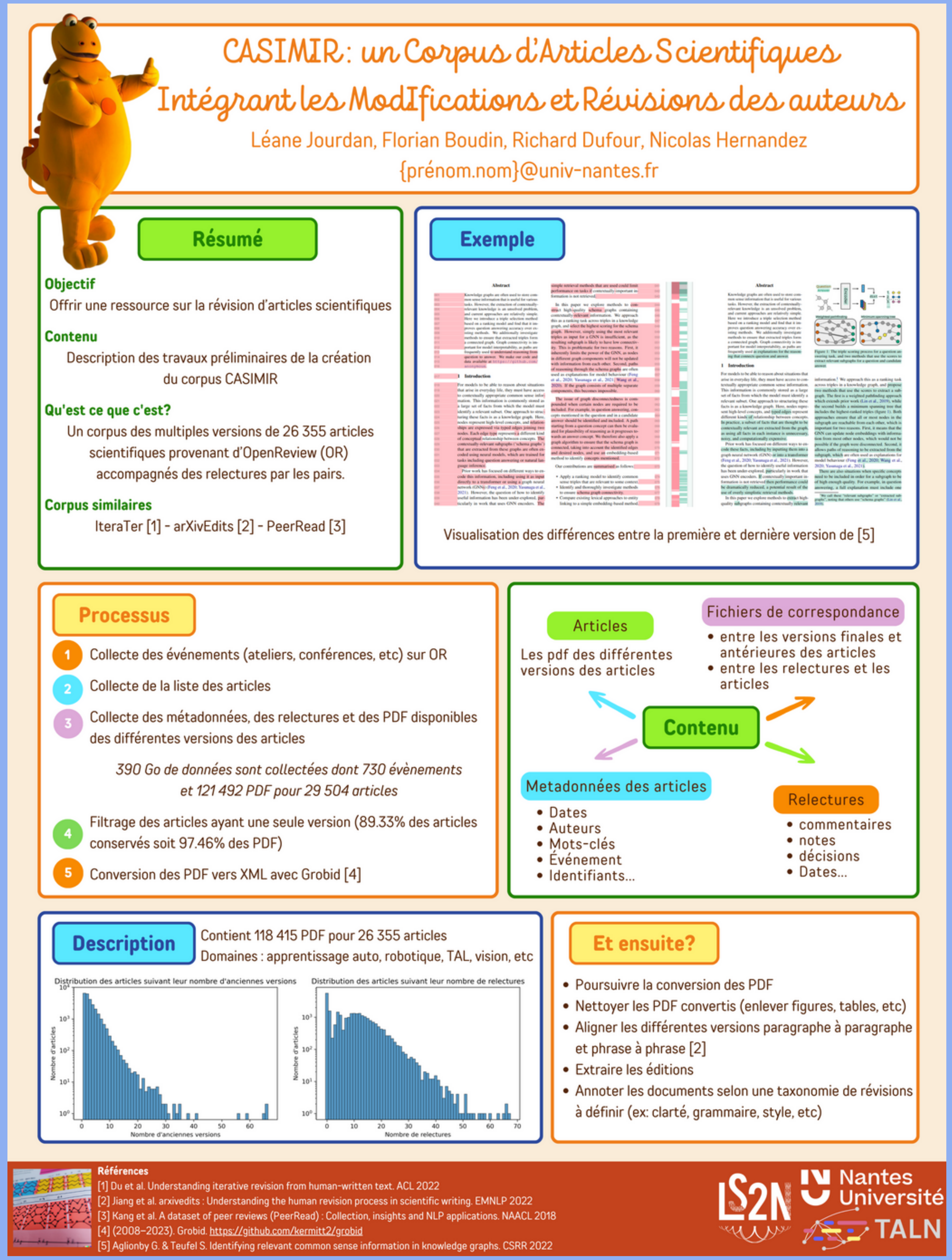
Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez
{prénom.nom}@univ-nantes.fr

Contenu de l'article

- Motivations
- Contenu du corpus
- Processus de collecte

Prochaines étapes et difficultés

- Conversion de PDF
- Alignement de documents
- Extraction de révisions
- Annotation des types de révisions



CASIMIR: un Corpus d'Articles Scientifiques Intégrant les ModIfications et Révisions des auteurs

Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez
{prénom.nom}@univ-nantes.fr

Résumé

Objectif
Offrir une ressource sur la révision d'articles scientifiques

Contenu
Description des travaux préliminaires de la création du corpus CASIMIR

Qu'est ce que c'est?
Un corpus des multiples versions de 26 355 articles scientifiques provenant d'OpenReview (OR) accompagnés des relectures par les pairs.

Corpus similaires
IteraTer [1] - arXivEdits [2] - PeerRead [3]

Exemple

Visualisation des différences entre la première et dernière version de [5]

Processus

1. Collecte des événements (ateliers, conférences, etc) sur OR
2. Collecte de la liste des articles
3. Collecte des métadonnées, des relectures et des PDF disponibles des différentes versions des articles

390 Go de données sont collectées dont 730 évènements et 121 492 PDF pour 29 504 articles

4. Filtrage des articles ayant une seule version (89.33% des articles conservés soit 97.46% des PDF)
5. Conversion des PDF vers XML avec Grobid [4]

Contenu

Articles
Les pdf des différentes versions des articles

Fichiers de correspondance
- entre les versions finales et antérieures des articles
- entre les relectures et les articles

Metadonnées des articles
- Dates
- Auteurs
- Mots-clés
- Événement
- Identifiants...

Relectures
- commentaires
- notes
- décisions
- Dates...

Description

Contient 118 415 PDF pour 26 355 articles
Domaines : apprentissage auto, robotique, TAL, vision, etc

Distribution des articles suivant leur nombre d'anciennes versions

Distribution des articles suivant leur nombre de relectures

Et ensuite?

- Poursuivre la conversion des PDF
- Nettoyer les PDF convertis (enlever figures, tables, etc)
- Aligner les différentes versions paragraphe à paragraphe et phrase à phrase [2]
- Extraire les éditions
- Annoter les documents selon une taxonomie de révisions à définir (ex: clarté, grammaire, style, etc)

Références
[1] Du et al. Understanding iterative revision from human-written text. ACL 2022
[2] Jiang et al. arxivedits : Understanding the human revision process in scientific writing. EMNLP 2022
[3] Kang et al. A dataset of peer reviews (PeerRead) : Collection, insights and NLP applications. NAACL 2018
[4] (2008–2023). Grobid. https://github.com/kermitt2/grobid
[5] Aglionby G. & Teufel S. Identifying relevant common sense information in knowledge graphs. CSRR 2022